



Introduction to Climate Data Services using iRODS

Data Management System Project

Savannah Strong Finch, Glenn Tamkin, Dave Ripley, Ed Luczak, Scott
Sinno, Roger Gill, Deni Nadeau, John Schnase, Mark Mcinerney

NASA Center for Climate Simulation (NCCS)

NASA Goddard Space Flight Center



Topics

- *NCCS Background*
- *Goals and Challenges*
- *Data Grid software (Overview)*
- *iRODS background*
- *Preliminary Tests with iRODS (Overview)*
- *Climate Data Server (Test Applications and Results)*
- *Current and Future Integration within NCCS (eCDS)*
- *Questions*



NCCS Background

Current

- Provide state-of-the-art high performance computing, storage, network, and application solutions to enable scientists to increase their understanding of the Earth and the universe
- Provide large-scale compute engines, analytics, data sharing, and high-end computing services support





Goals

Develop a data services capability to **better** support the climate research communities and prepare the way for technology advances for:

- *IPCC / AR5*
 - Provide the data management services and analytical tools necessary to support the publication requirements of the Intergovernmental Panel on Climate Change (IPCC).
- *Observation/Simulation Data Integration*
 - Bring the climate modeling and observational communities together to work toward the goal of integrating model outputs and observational data
- *Next Generation High End Computing (HEC) Requirements for Modeling and Assimilation*
 - Contribute emerging technologies to address computing requirements for Earth system modeling that will increase significantly in the coming years



Challenges

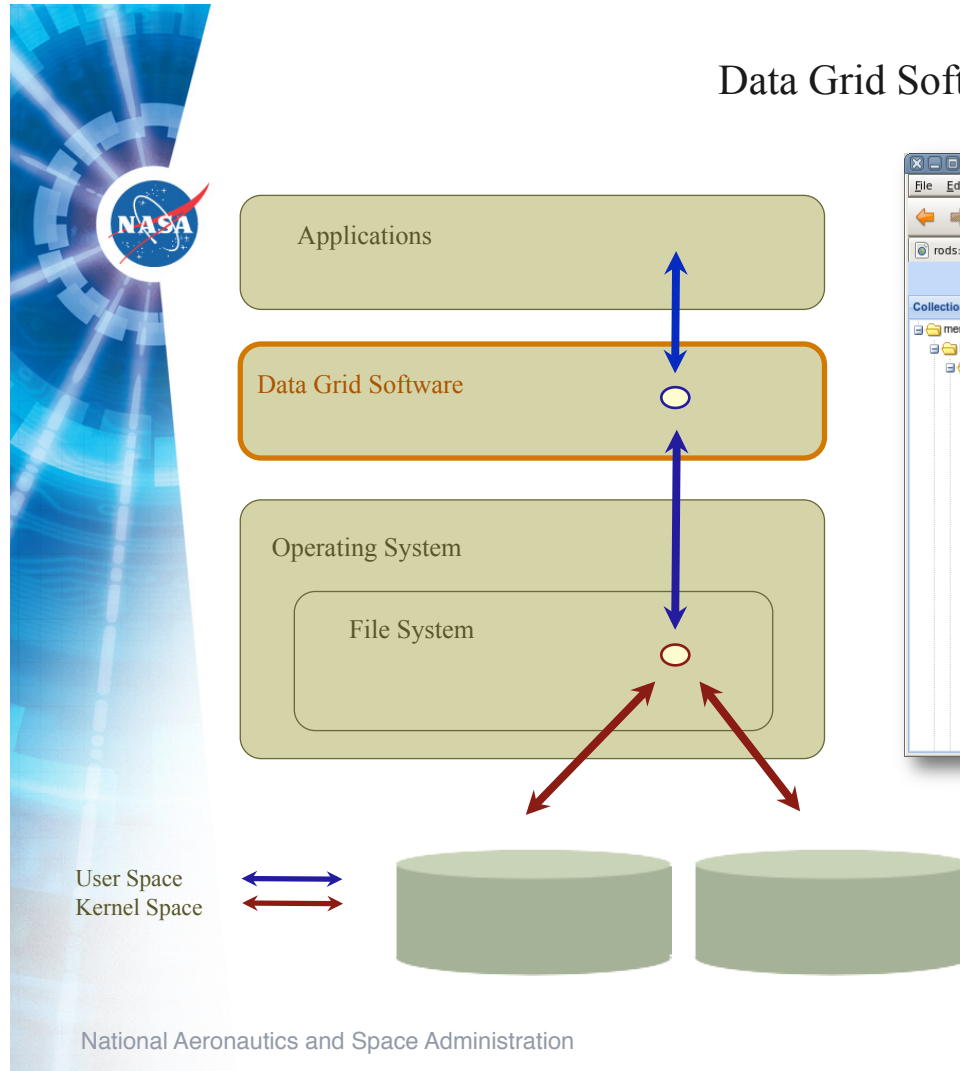
- *Finding* observational and model data for use in climate and weather studies
- *Accessing* the geographically distributed data
- *Managing* the massive digital holdings, which are measured in petabytes and hundreds of millions of files
- *Maintaining* the data, which must often be preserved for decades
- *Supporting* data sharing, data publication, and data stewardship





Data Grid Software

Data Grid Software



rods://rods@localhost:1247/merra_Zone/home/public/merra/1979 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

https://169.154.148.17/rods/browse.php#rurl=rods@localhost%3A1247/merra_Zone/home/public/mer...

rods://rods@localhost:1247/merra_Zo...

MERRA100.prod.assim.instM_3d_asm_Cp.197901.hdf

Sign Out

Open

Add Remove Reload Save

| Name | Value | Unit |
|---------------|---|------|
| variables | Sea-level pressure, Surface pressure, Surface Geopotential, Geopotential height, Ozone Mixing Ratio | |
| title | MERRA reanalysis. GEOS-5.2.0 | |
| source | Global Modeling and Assimilation Office. GEOSops_5_2_0 | |
| references | http://gmao.gsfc.nasa.gov/research/merra/ | |
| missing_value | 99999999+14f | |
| institution | Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD 20771 | |
| history | File written by CFIO | |
| hdfversion | HDFEOS_V2.14 | |
| dimensions | TIME:EOSGRID = 1, YDim:EOSGRID = 144, XDim:EOSGRID = 288, Height:EOSGRID = 42 | |
| conventions | CF-1.0 | |
| contact | http://gmao.gsfc.nasa.gov/ | |
| comment | GEOS-5.2.0 | |
| checksum | 91ddec7eee867abb8ca2e184ad2f8e92 | |

overview metadata Copies More

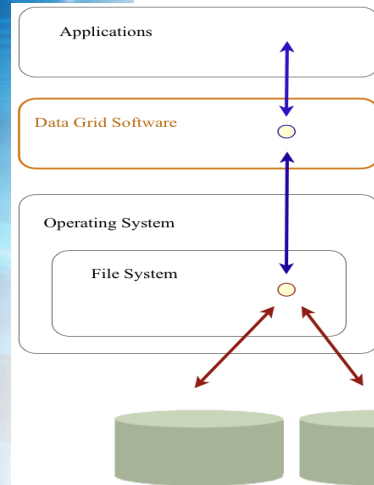
1997 Page 1 of 1

Displaying objects 1 - 12 of 12

Data grid “middleware” runs as an application in user space and provides a richer set of metadata descriptors and extended capabilities ...

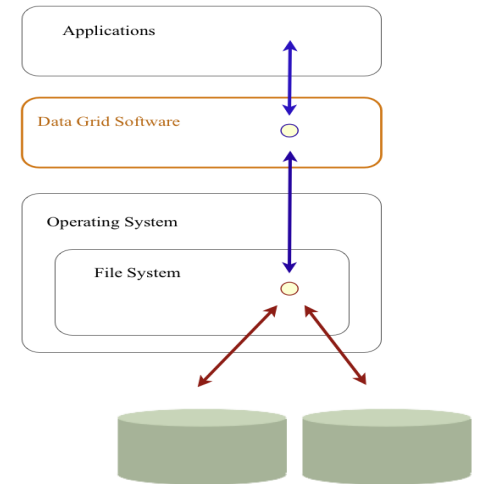
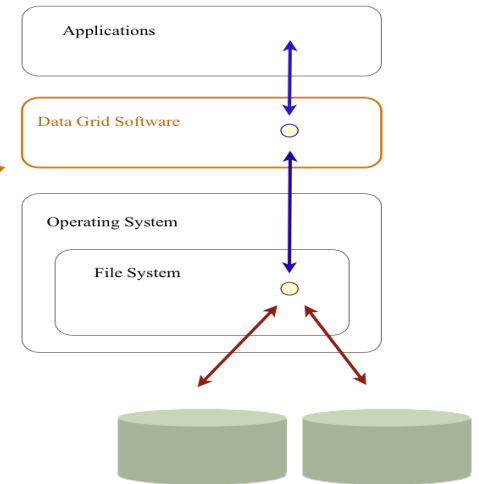
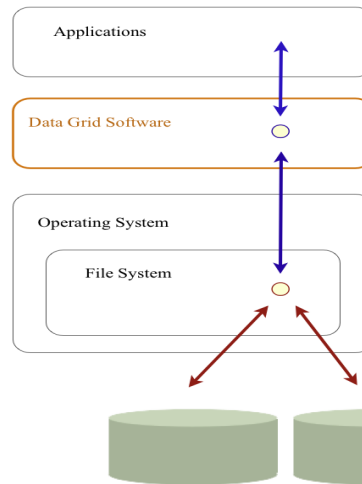


Data Grid Software



National Aeronautics and Space Administration

... including federation and
inter-collection discovery and access.





Data Grid Software

iRODS

Integrated Rule Oriented Data System

iRODS: integrated Rule-Oriented Data System

Background

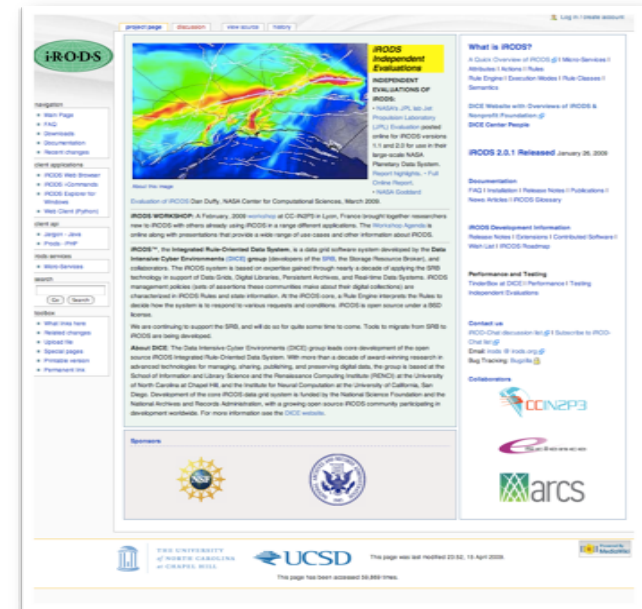
- Open source data grid software system.
- Developed by the Data Intensive Cyber Environments (DICE) group, University of North Carolina.
- Historic roots in data grids, digital libraries, persistent archives, and real-time data systems R&D, and SRB.

Features

- Management of large collections
- Manages metadata
- Policies, Rules and Micro-services
- A unified view of disparate data
- Controlled access
- Easy back up and replication
- High-performance network data transfer
- Support for a wide range of physical storage

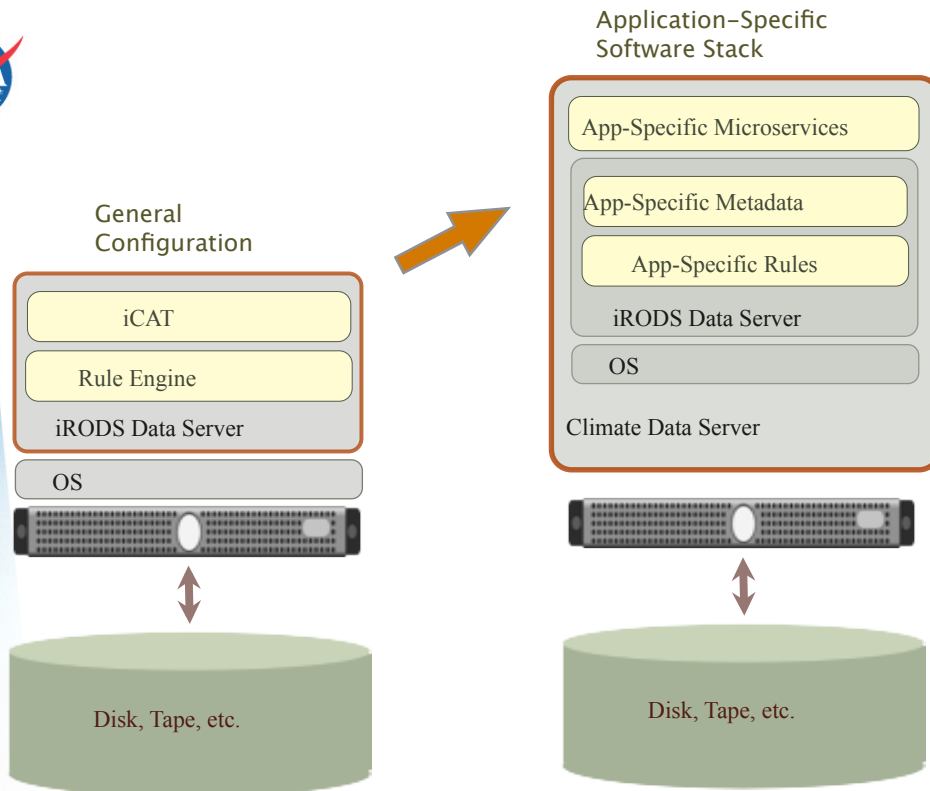
Major Concepts

- iRODS rules
 - Actions for policies
 - iRODS microservices.
 - Implementations of definitions of Actions
- *With iRODS metadata providing the information necessary to perform these mappings



www.irods.org

iRODS-Based Climate Data Server



Core Components

- Application-specific microservices
- Application-specific metadata
- Application-specific rules
- Application-specific utilities
- Application-specific configurations

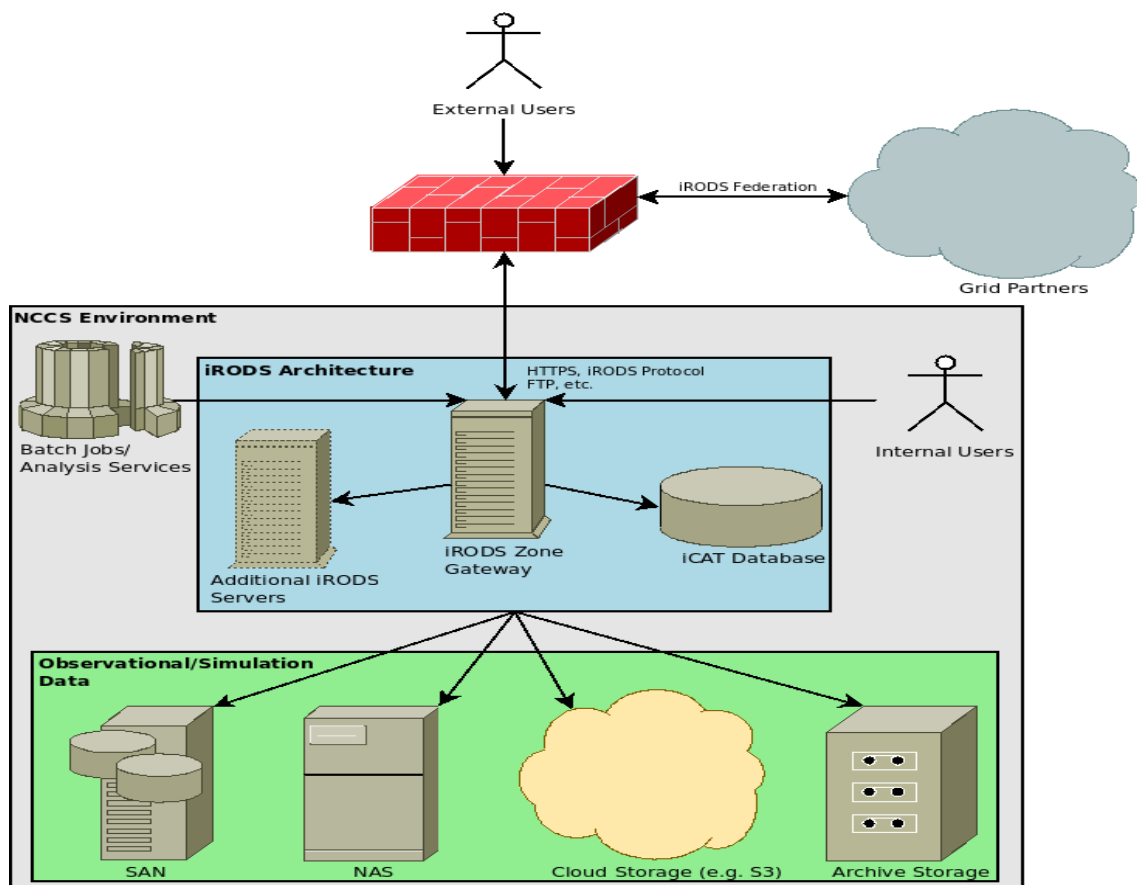
=> "Application-Specific Kit"

- A specific release of iRODS
- A specific operating system

=> "CDS Software Appliance"

iRODS ..

- iRODS abstracts physical location of data
- iRODS assists with archive management



Preliminary Tests – Ingest/Registration/View

iRODS rules and microservices allow data to be stored in configurable collections based on data policies

- Rich web client allows for “explorer” like view into collections of the registered data
- Can also perform command line interface “icommands”:
 Bash-4.1\$ ls
 /merra_Zone/home/public/merra/1979:
 MERRA100.prod.assim.instM_3d.nc
 .
 .

*Replication to backup storage resources also supported

The screenshot shows the iRODS web client interface in a Mozilla Firefox browser. The address bar displays the URL `https://169.154.148.17/irods/browse.php#ruri=rods@localhost%3A1247/merra_Zone/home/public/merra/1979`. The interface features a sidebar on the left with a tree view of collections, including `merra_Zone`, `home`, `public`, and `merra`. The `merra` collection is expanded, showing a list of years from 1979 to 2009. The main panel displays a table of files within the `1979` collection. The table has columns for Name, Resource, Size, and Date Modified. The files listed are all named `MERRA100.prod.assim.instM_3d.nc` and are stored in the `demoResc` resource. The sizes range from 125.85 MB to 128.45 MB, and the dates are all from July 14, 2010.

| Name | Resource | Size | Date Modified |
|---------------------------------|----------|-----------|-------------------------|
| MERRA100.prod.assim.instM_3d.nc | demoResc | 125.88 MB | July 14, 2010, 11:02 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 126.24 MB | July 14, 2010, 11:07 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 127.29 MB | July 14, 2010, 11:11 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 128.23 MB | July 14, 2010, 11:15 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 126.73 MB | July 14, 2010, 11:19 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 126.46 MB | July 14, 2010, 11:24 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 124.78 MB | July 14, 2010, 11:29 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 125.34 MB | July 14, 2010, 11:33 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 126.65 MB | July 14, 2010, 11:37 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 128.45 MB | July 14, 2010, 11:41 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 127.45 MB | July 14, 2010, 11:46 am |
| MERRA100.prod.assim.instM_3d.nc | demoResc | 125.85 MB | July 14, 2010, 11:50 am |



Preliminary Tests - Search

- iRODS rules and microservices can be used to assign metadata
- iRODS provides advanced search capabilities over the metadata

The screenshot displays the iRODS web interface in a Mozilla Firefox browser. The address bar shows the URL: `https://169.154.148.17/irods/browse.php?rurl=rods%3Alocalhost%3A1247/merra_Zone/home/public/merra/1979`. The interface includes a sidebar with a file tree showing collections like 'merra_Zone', 'home', and 'public'. An 'Advanced Search' dialog box is open, showing search criteria: Name (Name or Partial Name, case sensitive), Modified Within (Any Time), Owner (Owner of the file), Resource (Resource of the file), and Current Collection (/merra_Zone/home/public/merra/1979). Below the search criteria, a 'Metadata' section shows a table with columns for variable, operator, and value. The table contains the following data:

| variables | like | Surface Geopotential |
|-----------|------|-----------------------|
| checksum | = | 67abb8ca2e184ad2f9e92 |
| Name | Op | Value |
| Name | Op | Value |
| Name | Op | Value |

At the bottom, a 'Metadata' tab is selected, showing a table with the following data:

| Metadata | Value |
|---------------|---|
| hdfEOSversion | HDFEOS_V2.14 |
| dimensions | TIME:EOSGRID = 1, YDim:EOSGRID = 144, XDim:EOSGRID = 288, Height:EOSGRID = 42 |
| conventions | CF-1.0 |
| contact | http://gmao.gsfc.nasa.gov/ |
| comment | GEOS-5.2.0 |
| checksum | 91ddec7eee867abb8ca2e184ad2f9e92 |




Climate Data Server (Test Applications)


- **MODIS and ISDS**
- **Merra Monthly Means and Proxies for AR5 simulations**
- **vCDS in the Amazon Cloud**
- **ODAS workflow**

Climate Data Server (Test Applications) – Observational Data

- Developed an iRODS data grid that published Moderate Resolution Imaging Spectroradiometer (MODIS) observational data
 - 54 million registered files, 630 TB of data, and over 300 million defined metadata values
- Developed an iRODS data grid that focuses on a small-scale, multi-product, application-specific data service
 - The Invasive Species Data Service (ISDS) manages a collection of MODIS data products for ecological forecasting applications




modis_Zone




| | |
|--------------------|---|
| Collection | MODIS Atmosphere |
| Data | Aerosol, Water Vapor, Cloud, Profile, Cloud Mask, Joint Products |
| Type | Observational |
| Format | HDF |
| Customers | GES DISC, MODIS Science Data Support Team / Admins, Users |
| Distinction | Operational environment, 40,000,000 Data Objects! |
| Interfaces | Programmatic (Admin), FUSE (User), iRODS clients |
| Status | TRL 3 => TRL 6 <i>(Subsystem validation in an operational environment.)</i> |

One of the most important ecological issues concerning our planet is climate change. It is generally agreed that the Earth's climate will modify in response to radiative forcing induced by changes in atmospheric trace gases, cloud cover, cloud type, solar radiation, and climate change.

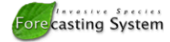



isds_Zone




| | |
|--------------------|---|
| Collection | Invasive Species Data Service (ISDS) |
| Data | MODIS Land NDVI Phenology (Time-Series) Data |
| Type | Observational |
| Format | GeoTIFF |
| Customers | MD DNR, DOI BLM (GSENM) / Users |
| Distinction | Personal-/laboratory-scale, application-specific (ISFS) iRODS-based DMS; ISFS/ISDS licensed for distribution |
| Interfaces | isds_CI (User), iRODS clients |
| Status | TRL 3 => TRL 7 <i>(System validation in an operational environment.)</i> |

The Invasive Species Forecasting System (ISFS) is a modeling framework that allows users to load point occurrence field sample data for a plant species of interest and quickly generate habitat suitability maps for geographic regions of management concern, such as a national park, monument, forest, or refuge. Target customers for applications built using ISFS are natural resource managers and decision-makers who have a need for scientifically valid, model-based predictions of the habitat suitability of plant species of management concern.





Climate Data Server (Test Applications) – Analysis and Simulation Data


- Developed an iRODS data grid that manages Modern Era Retrospective-Analysis for Research and Applications MERRA Monthly means analysis data
 - 360 files, 47 GB of data, and 4000 metadata values
- Developed an iRODS data grid that published public GEOS-5 simulation data as a proxy for AR5 data sets
 - 134,000 files, 12 TB of data, and 400,000 metadata values




merra_Zone



| | |
|--------------------|---|
| Collection | Modern Era Retrospective-Analysis for Research and Applications (MERRA) |
| Data | Monthly products from the past 15 years |
| Type | Observational/Simulation |
| Format | NETCDF |
| Customers | NCCS, GES DISC, ESG, Nebula / Admins, Managers |
| Distinction | merra_Zone @ NCCS + merra_Zone @ Nebula |
| Interfaces | merra_CI (Admin), iRODS clients |
| Status | TRL 3 => TRL 5 (System validation in a relevant test environment) |




yotc_Zone



| | |
|--------------------|---|
| Collection | Year of Tropical Convection (YOTC) |
| Data | Satellite, in-situ and simulation/prediction model data sets |
| Type | Observational/Simulation |
| Format | NETCDF |
| Customers | NCCS / Admins, Users |
| Distinction | Operational environment, iRODS-mediated archive management |
| Interfaces | yotc_CI (Admin), FUSE (User), iRODS clients |
| Status | TRL 3 => TRL 7 (System validation in an operational environment.) |

The realistic representation of tropical convection in our global atmospheric models is a long-standing grand challenge for numerical weather forecasts and global climate predictions. To address the challenge of tropical convection, collaborative organizations from around the world have proposed a year of coordinated observing, modeling and forecasting of organized tropical convection and its influences on predictability. This effort is intended to exploit the vast amounts of existing and emerging observations, the expanding computational resources and the development of new, high-resolution modeling frameworks, with the objective of advancing the characterization, diagnosis, modeling, parameterization and prediction of multi-scale convective/dynamic interactions, including the two-way interaction between tropical and extra-tropical weather/climate.

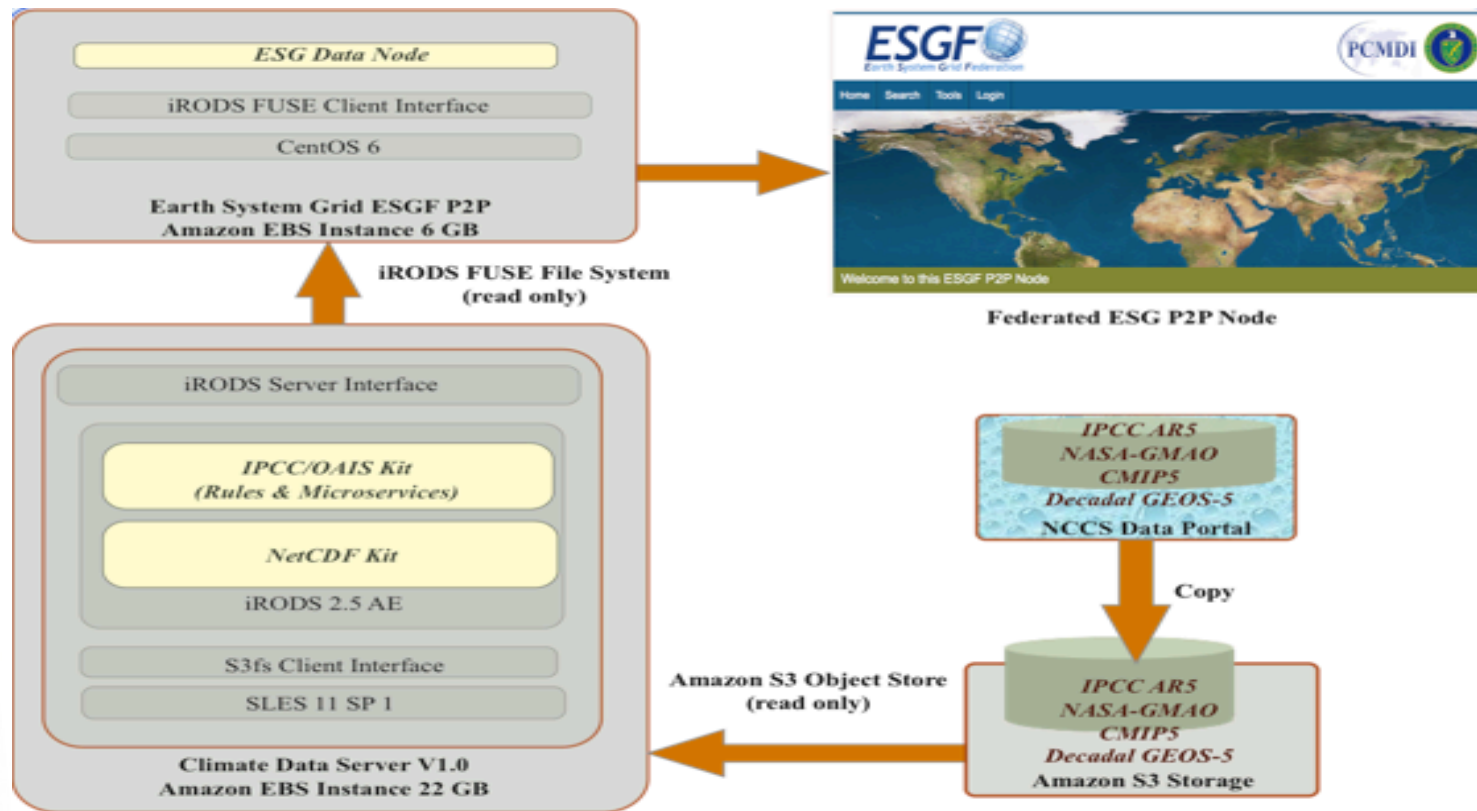




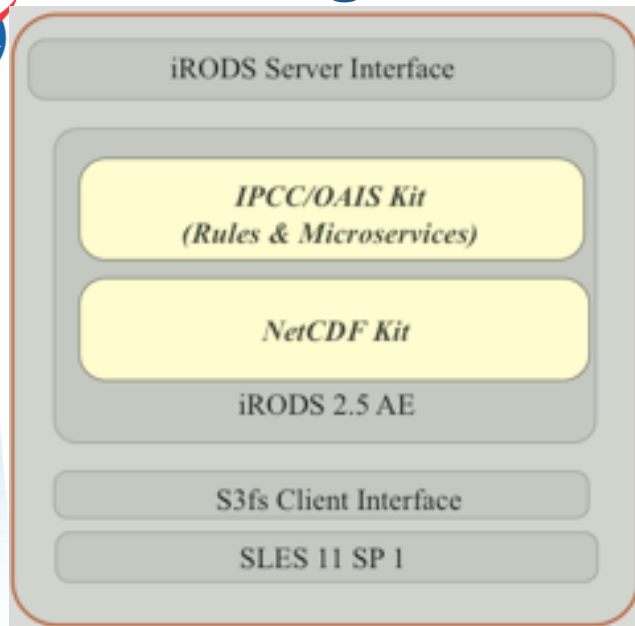
Climate Data Server (Test Application) – Federation

- Tested and evaluated iRODS data federation
 - Federated the GEOS-5 public data and MODIS grids to simulate the union of observational and simulation data
- the integrated management of observational and simulation data was explored
 - Implemented an interface that enables comingling of remote and local observational and simulation data for advanced scientific study

Climate Data Server (Test Application) – vCDS in the Amazon Cloud



Climate Data Server (Test Application) – Extending iRODS



- NetCDF kit knows how to read the file header based on file format
- IPCC/OAIS kit defines which metadata to store and how to store it

A006.ocn_ana_2D.19950101.nc

Open

Add Remove Reload Save

| Name | Value | Unit |
|-----------------------|---------------------------------------|------|
| var:VS:_FillValue | 1.e+15f | |
| var:VS:vmin | -1.e+15f | |
| var:VS:vmax | 1.e+15f | |
| var:VS:valid_range | -1.e+15f, 1.e+15f | |
| var:VS:units | m s-1 | |
| var:VS:type | float | |
| var:VS:standard_name | surface_Agrid_northward_velocity | |
| var:VS:scale_factor | 1.f | |
| var:VS:missing_value | 1.e+15f | |
| var:VS:long_name | surface_Agrid_northward_velocity | |
| var:VS:fmissing_value | 1.e+15f | |
| var:VS:dim | time, lat, lon | |
| var:VS:coordinates | LON LAT | |
| var:VS:add_offset | 0.f | |
| var:VI:_FillValue | 1.e+15f | |
| var:VI:vmin | -1.e+15f | |
| var:VI:vmax | 1.e+15f | |
| var:VI:valid_range | -1.e+15f, 1.e+15f | |
| var:VI:units | m s-1 | |
| var:VI:type | float | |
| var:VI:standard_name | meridional velocity of surface seaice | |

overview metadata Copies More

Climate Data Server – Ocean Data Assimilation System (ODAS)

- Leveraged iRODS to monitor ODAS workflow status
- Developed a series of “Ocommands” that are wrappers around the iRODS “icommmands”
- Ocommands were integrated into existing ODAS workflow scripts and perform functions such as:
 - Register data
 - Query the iRODS database for decision-making information
 - Maintain the status of the hierarchy of ODAS workflow artifacts as status changing events occur
 - Log relevant event metadata to the appropriate ODAS workflow artifact
 - Reprocess preparation
 - Remove relevant files and reset status in the hierarchy of ODAS workflow artifacts in preparation for reprocessing.

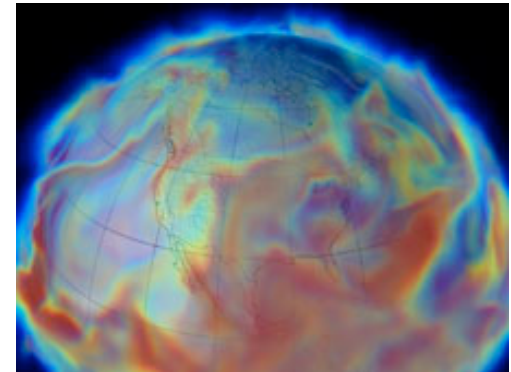
GEOS5odas-5.00_odas-503_B001_exp2010.xml

Open

| Name | Value | Unit |
|--------------------|---|------|
| @year | 2010 | |
| @id | 20263da2-645e-4308-9fdf-b592f98313ec | |
| @tag | 503 | |
| @expname | B001 | |
| @category | EXPERIMENT | |
| @odas | GEOS5odas-5.00 | |
| @dataflow_id | GEOS5odas-5.00_odas-503:B001 | |
| @date_created | Mon Feb 4 12:29:57 EST 2013 | |
| @exception | n/a | |
| @last_processed | n/a | |
| @status_exp | PENDING | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201001.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201002.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201003.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201004.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201005.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201006.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201007.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201008.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201009.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201010.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201011.xml | |
| @status_detail_run | PENDING: GEOS5odas-5.00_odas-503_B001_run201012.xml | |
| @user_id | rip3 | |
| @type | XML | |
| @description | Yearly processing rollout for 2010 | |

Results

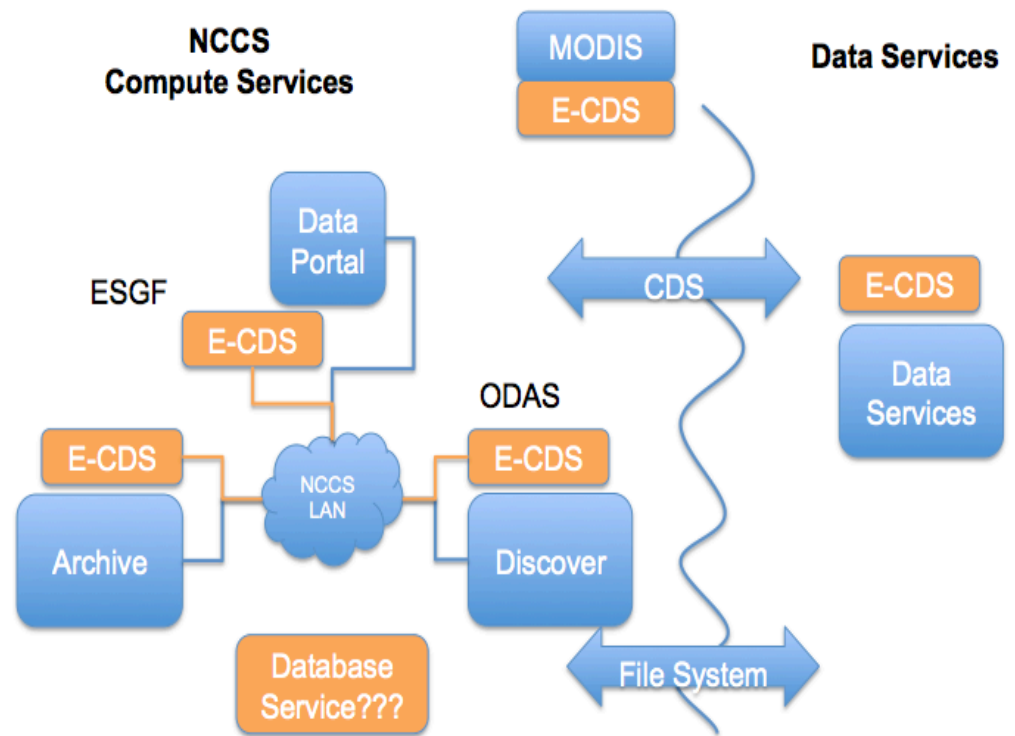
- iRODS is a promising technology for exposing services for data management, publication, and analysis
- The iRODS catalog (ICAT) demonstrated adequate scaling for data registration
 - Optimization desired for searching huge datasets
- Good collaboration with the iRODS development team
- Exercised enough diverse Test Cases to have confidence in performance leading to decision to be made to progress towards making iRODS-based Climate Data Services Operational





Moving Forward - Enterprise Climate Data Services E-CDS

- Establish an Enterprise Climate Data Service (E-CDS) federated grid across the NCCS resources
- Starting with projects:
 - ODAS
 - ESGF
 - Archive
 - Allows for operational capability of the ESGF to use the archive in the case that the data portal disks were unavailable
- Potential follow on projects MERRA2, NCA, UVCDAT, MODIS





The End.



Questions?



Moving Forward – What does E-CDS mean to operations folks

- Account creation
- Config (firewall, security, etc..)
- e-CDS Dependency installation (unixODBC, postgres, perl, authd, etc..) and configuration
- Installation of E-CDS rpm (includes irods + extensions)
- Admin Support